Journal Title
XX(X):1–24

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A new parsimonious model for ordinal longitudinal data with application to subjective evaluations of a gastrointestinal disease

Moreno Ursino[1,2] and Mauro Gasparini[1]

Abstract

In this paper a new discrete statistical model for ordered categorical data is proposed via fixed-point discretization of a beta latent variable. The resulting discretized beta distribution has a highly flexible shape and it can be either over-dispersed or under-dispersed with respect to the binomial distribution. It has only two parameters, which may therefore parsimoniously depend on covariates and on random effects, providing new tools for the analysis of structured, clustered or longitudinal ordinal data. Practical examples and advices are given and an application of the new model to subjective evaluations of a gastrointestinal disease is shown.

Keywords

Longitudinal data; Mixed effect model; Ordinal data; Ordinal regression, Beta distribution

[1]INSERM, UMRS 1138, team 22, France.
[2]Department of Mathematical Sciences, Politecnico di Torino, Italy.

Corresponding author:
Moreno Ursino, INSERM, UMRS 1138, team 22, CRC, University Paris 5, University Paris 6, 15 Rue de l'École de Mèdecine, 75006 Paris, France.
Email: moreno.ursino@inserm.fr

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

1 Introduction: ordinal longitudinal data

Ordered categorical (ordinal) data are commonplace in the medical and health sciences. In order to diagnose or classify patients, it is usual to assign them into ordered categories corresponding to various degrees of a medical condition or levels of risk of developing a disease. For our purposes, the defining property of ordinal data is that there exist a clear ordering of the response categories, but no clear underlying interval scale between them. Methods developed for categorical data in general can be applied to the analysis of ordered categorical data, resulting as a consequence in some loss of information. There are important advantages in using models and methods developed explicitly to take into account the order of the categories. In particular, models for ordered categorical data tend to be more parsimonious than their unordered counterparts, thus resulting in more efficient inferences which facilitate the interpretation of parameters.

The methodology for ordinal data focuses on two general areas of statistical analysis, *association* and *regression*, the main references in the area being Agresti[1] and Becker[2]. The main focus of this work is on regression models. Logistic regression and loglinear models for categorical data were primarily developed in the 1960s and 1970s. Although ordinal data received some attention in those years – see e.g. Snell[3] and Bock and Jones[4] – a stronger focus was inspired later by articles such as Cullagh[5] on logit modeling of cumulative probabilities and Goodman[6] on loglinear modeling of the odds ratios. There are multiple approaches to defining logits for the response, but the three types that figure most prominently in the biostatistics literature are the adjacent-categories logits, the continuation-ratio logits, and the cumulative logits by Cullagh[5], the most popular one. We indicate such cumulative logit model as CLM from now on.

Alternative approaches to modelling ordinal data without reference to odds ratios or their logarithms center on using distributions on the first few integers $0, 1, \ldots, n$, interpreted simply as codes for the ordered categories. Within these approaches, it is important that the metric properties of $1, \ldots, n$ do not play a dominant role in the interpretation of the results, since $1, \ldots, n$ are simply convenient labels for the ordered levels. These approaches range from the use of a simple binomial distribution to more flexible distributions which allow for overdispersion, such as the beta-binomial (see for example Muniz-Terrera et al.[7], in which GAMLSS models[8] are used to analyse cognitive test data), to the more recent approach of D'Elia and Piccolo[9], who proposed to use a mixture – called CUB or MUB – of a binomial and a discrete uniform distribution. The idea of the latter authors, who take a psychometric point of view, is that the response is the result of a combination of *feeling* and *uncertainty* components which can be modeled parsimoniously using only two parameters, obtaining nonetheless various possible shapes and behaviours of the response distribution. Several extensions of CUB models are presented in more recent work (see for example Iannario[10]).

A more general area, in which there has been a lot of recent activity, is the analysis of *repeated measures* data in the form of ordered categorical responses. Such data arise, for example, in longitudinal studies, crossover experiments, studies of families or kinship and they all exhibit some sort of clustering. The analysis of repeated measures data have been applied to these problems along three main approaches: *conditional*, *marginal* and *transition* models.

In this work we focus on conditional models, also called *subject-specific models*, which describe each cluster individually, making the mean response dependent not only on covariates but also on a vector of random effects associated with the cluster. Mixed effects models focus initially on the regression relationship restricted to observations on a single cluster (or individual, like a patient). The model is then extended to several individuals by allowing some features of the model – the random effects – to vary from individual to individual in a prescribed manner, while other components – the fixed effects – remain the same. This approach describes and interprets the covariance structure for longitudinal observations only indirectly, as opposed to other approaches focussing on the error correlations. When the vector of random effects contains more than just a few terms, model fitting via maximum likelihood estimation can be challenging. The classical approach is based on maximizing the marginal likelihood obtained by integrating out the random effects from the joint likelihood as shown in Lee and Daniels[11]. Except in rare cases[12;13], this integral does not have a closed form and some approximations are necessary. Pseudo-likelihood methods can avoid the problems associated with integration, making computations simpler[14]. Unfortunately, these methods are biased for highly non-normal distributions like those used with binary or ordinal data[14;15], with the bias increasing as variance components increase. Recently, Bayesian estimation of mixed-effect model arose as an alternative to MLE since Bayesian hierarchical models naturally fit the mixed-effect structure[16;17]. Posterior distributions are computed directly from the joint likelihood and, at the end of the process, all random effects are estimated. Integrations are left to the optimization algorithm and, using priors, Bayesian methods can deal to smaller sample sizes. An example of Bayesian estimation applied to longitudinal data and based on the cumulative logit assumption can be found in Lunn et al[18]. On the other hand, computational times are usually longer than the MLE counterpart.

Marginal models, also called models with *population-averaged* effects, refer to averaging over clusters: the mean response depends only on the covariates of interest, and not on random effects or previous responses. Maximum likelihood estimation for these models is computationally challenging, and generalized estimating equations (GEEs) are often used[19].

Finally, transition models describe the distribution of a response conditional on past responses and explanatory variables. This approach has received substantial attention for binary data[20]. Many transitional models have a Markov chain structure to describe time evolution[21].

The choice among marginal, conditional and transitional models depends on whether one prefers to interpret data at the population or at the subject level or whether it is sensible to describe effects of

explanatory variables conditional on previous responses. Conditional models are especially useful for describe within-cluster effects, such as within-subject comparisons in a crossover study.

To recap, we present a new model to analyse ordinal data and extend it to the longitudinal case following the conditional approach. In Section 2 a motivating case study regarding gastroesophageal reflux disease is described. In Section 3 our discretized beta model is introduced. We postulate the existence of a beta latent random variable and discretize it using fixed cut-points. In this way, we find a discrete probability distribution which fits an ordinal response variable Y in a parsimonious way and lends itself to a manageable treatment in the presence of covariates and random effects. Section 4 concerns parameter estimation for a single population and in presence of covariates. In Section 5 an extension to longitudinal data is presented where, taking into account the correlation between the repeated responses, random effects are added. Both Maximum Likelihood and Bayesian approaches are discussed. Section 6 gives some practical considerations on the new distribution. Several scenarios and situations are discussed in order to point out the strengths and the limits of this regression. Finally, a real data set consisting of subjective evaluations of a gastrointestinal disease is analysed.

2 Motivating case study

The primary aim of this study was to assess the utility of the water load test (WLT) as a tool to classify patients with gastroesophageal disease. The WLT is a non-invasive, inexpensive, easily performed drink test, well tolerated and reproducible in healthy subjects and in patients with functional dyspepsia or gastroesophageal reflux disease (GERD). Battaglia et al.[22] show that in GERD patients, with mild erosive esophagitis and non erosive reflux disease, the WLT is abnormal, similar but non identical to that reported in patients with functional dyspepsia. In this study, we wanted to investigate if there is any difference in gastric sensations between healthy volunteers and GERD patients as assessed by the WLT.

71 GERD patients and 30 healthy volunteers with no abdominal symptoms or history of upper gastrointestinal disorders were recruited. None of the subjects had previously undergone abdominal surgery, except for appendectomy. The study was carried out according to local ethical rules, with informed consent, and in accordance with the recommendations of the Helsinki Declaration.

The WLT was performed on all subjects by giving them room temperature water for 5 minutes or until they perceiving the “full” stomach sensation. Total volume consumed (ml) was determined and recorded.

Just after finished drinking (time=0) and then after 30 minutes (time=1), participants completed a symptom visual analogue scale (VAS; integer from 0, absent, to 10, maximal) to score nausea, postprandial fullness and bloating. The resulting data is a first example of longitudinal data in the sense that all subjects produced data at time 0 and time 1. Actually, this data over two times was just the

beginning segment of more complex data set which would allow for a diversified therapy for each patient, but for the sake of simplicity we consider here only these first two times.

3 The discretized beta distribution

In the now classical approach to ordinal data, a normal or a logistic distribution is often assumed as the underlying latent variable[23]. The normal distribution has two unknown parameters to be estimated, to which a number of cut-points equal to the number of levels of the ordinal variable minus 1 are added. This fairly large number of parameters may be hard to estimate and a more parsimonious approach using fewer parameters may be preferred, especially when clustered repeated measures are taken. This is what led D'Elia and Piccolo[9] to propose their CUB model, based on two parameters but still providing a variety of possible probability distributions over the first few integers, used as codes for ordinal variables. We present an alternative approach which, like the classical approaches to ordinal data, assumes an underlying latent variable and, like D'Elia and Piccolo, reduces drastically the number of parameters. We choose the beta distribution for the latent variable, which has compact support and a variety of possible bell-shapes, U-shapes and J-shapes, and discretize it using fixed cut points.

More precisely, we assume that Y^*, the latent variable, follows a beta distribution with density

$$f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0,1) \tag{1}$$

where $\mu \in (0,1)$, $\phi > 0$ and $\Gamma(\cdot)$ is the gamma function $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,dx$. This is not the usual parametrization of the beta law (which is instead often parametrized by $\alpha = \mu\phi$ and $\beta = (1-\mu)\phi$), but it is convenient for modeling purposes and it is often used in so-called Beta-regression models[24]. Since $E(Y^*) = \mu$ and $\text{Var}(Y^*) = \mu(1-\mu)/(1+\phi)$, we can see that μ is really the mean of the latent variable and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ, the larger the value of ϕ, the smaller the variance of Y^*. To define our new distribution, we discretize Y^* using n fixed cut points, with n equal to the number of ordinal levels minus one. In compact form,

$$Y = \lfloor (n+1)Y^* \rfloor,$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Notice that in this way the cut points are known constants, not parameters. In more detail, we define an integer-valued random variable Y such that $Y = k$ if and only if $Y^* \in \left[\frac{k}{n+1}, \frac{k+1}{n+1}\right)$. In other words, the probability that Y is equal to k is the same as the probability that the latent variable Y^* falls in the interval $\left[\frac{k}{n+1}, \frac{k+1}{n+1}\right)$:

$$P(Y=k) = \int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} f(x)\,dx = I_{\frac{k+1}{n+1}}(\mu\phi,(1-\mu)\phi) - I_{\frac{k}{n+1}}(\mu\phi,(1-\mu)\phi) \qquad k=0,1,\ldots,n \quad (2)$$

where $I_x(p,q)$ is called the *incomplete beta function ratio*[25], called simply *incomplete beta function* from now on, and it is defined as $I_x(p,q) = B(p,q)^{-1}\int_0^x t^{p-1}(1-t)^{q-1}\ dt$ with $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\ dt = \Gamma(p)\Gamma(q)/\Gamma(p+q)$. We then say that Y follows a *discretized beta distribution* and write

$$Y \sim \text{Dbeta}(\mu,\phi,n)$$

. Note that our distribution is different from the one defined by Punzo[26].

The main properties of the discretized beta distribution are:

- for $n > 2$ the discretized beta family is identifiable;
- the discretized beta distribution can be either over-dispersed or under-dispersed with respect to the binomial distribution (unlike, for example, the beta-binomial distribution, which is always overdispersed);
- the uniform distribution is a special case of the discretized beta distribution;
- binomial distributions as well as beta-binomial distributions (often used as an alternative to binomial distributions with overdispersed data) can be approximated well by discretized beta distributions.

Proofs are in the Appendix.

Our approach is somewhat opposite to Johnson and Albert[23]: their latent distribution has a fixed form, e.g. a logistic with known variance, and the value of the mean of such distribution together with the estimated cut-points give us the probabilities of the discrete random variable Y; in our approach, on the contrary, we have the flexible shape of the latent variable Y^* defining the probabilities of Y, whereas cut points are fixed.

The uniform subdivision of the finite support of the beta distribution gives Y the same trend of the latent variable Y^*. This way, the usual scenarios encountered in the study of ordinal outcomes are well described by uni-modal beta latent variables, whereas in rarer, more extreme, cases bimodal shapes are necessary and are then provided by U-shaped latent distributions. For example, in survey data in which a respondent is asked to characterize his opinions on a scale ranging from "strongly disagree" to "strongly agree", he may be doubtful between two or three adjacent values, but rarely he may be doubtful between two distant values, such as "strongly disagree" and "strongly agree". Similarly, in the medical field, when a patient is asked to quantify his pain with a symptom visual analogue scale (VAS) ranging on the integers from 0 to 10, he may be doubtful between 5, 6 or 7, but rarely between 1 or 9. Thus, it often makes sense

to use unimodal distributions. On the other hand, the precision parameter ϕ quantifies how strongly this opinion is focused on just one or just few adjacent values.

4 Fitting discretized beta distributions

In this section, let $Y_1, \ldots, Y_N$ be a sample of independent observations each having a discretized beta distribution; in the next subsection we consider the i.i.d. case and assume that individual responses follow the same discretized beta distribution, with fixed unknown parameters to be estimated, then we move to the case in which they can have different parameters. In both situations, we use maximum likelihood to estimate unknown parameters.

4.1 The i.i.d. case

In the i.i.d. case, let $\mathbf{y} = (y_1, ..., y_N)$ be is the vector of the observed values; the likelihood is then

$$\mathcal{L}(\mu, \phi | \, \mathbf{y}) = \prod_{k=0}^{n} \left[I_{\frac{k+1}{n+1}} \left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}} \left(\mu\phi, (1-\mu)\phi\right) \right]^{n_k}, \tag{3}$$

where n_k is the absolute sample frequency of the $k-$th ordinal level. The maximum likelihood estimate (MLE) of the parameter $\boldsymbol{\theta} = [\mu, \phi]$, is computed as the maximizing value

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \sum_{k=0}^{n} n_k \log \left[I_{\frac{k+1}{n+1}} \left(\mu\phi, (1-\mu)\phi\right) - I_{\frac{k}{n+1}} \left(\mu\phi, (1-\mu)\phi\right) \right]. \tag{4}$$

Due to the parameter constraints, we can use the "L-BFGS-B" method, a quasi-Newton method[27], implemented in the `optim` function in the `R` software. A good starting point for μ and ϕ is given by $(\mu_0, \phi_0) = \left(\frac{1}{n+1} \frac{\sum_i y_i}{N}, (n+1) \frac{N-1}{\sum_i (y_i - \mu_0)^2} \right)$, since μ is related to the mean end ϕ to the inverse of the variance.

4.2 Independent observations with predictors

In this section suppose $(Y_1, ..., Y_N)$ are independent observations and a vector of covariates $\mathbf{x}_i$ (predictors) is also observed on the $i-$th observation. We can not follow the approach of generalized linear models, in which the link function of the mean of response variable is allowed to vary linearly with the predictor values, rather than assuming that the mean response itself vary linearly. In our case, if Y follows a discretized beta distribution for a given n, several pairs of different (μ, ϕ) generate the same expectation, as shown in Appendix A.2. We therefore introduce covariates on μ and ϕ without direct

reference to the expectation of Y. In order to specify a correspondence between the values of covariates and the parameters μ and ϕ, we propose the logit link for μ and the log link for ϕ, i.e.

$$\begin{aligned} \text{logit}\,(\mu_i) &= \mathbf{x}_i^{\mu}\boldsymbol{\beta}; \\ \log\,(\phi_i) &= \mathbf{x}_i^{\phi}\boldsymbol{\gamma}; \\ i &= 1, \cdots, N, \end{aligned} \tag{5}$$

where μ_i and ϕ_i are the parameters of the ith individual, $\mathbf{x}_i^{\mu}$ and $\mathbf{x}_i^{\phi}$ are subvectors of $\mathbf{x}_i$, possibly different but including usually the intercept, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \cdots, \gamma_{q-1})$, p and q being the lengths of $\mathbf{x}_i^{\mu}$ and $\mathbf{x}_i^{\phi}$ respectively.

For example, let Y_i be the ith patient's response on a VAS for the nausea symptom, let $\mathbf{x}_i^{\mu}$ be the vector composed of the intercept, the values of age, an indicator variable for sex and BMI (body mass index) and let $\mathbf{x}_i^{\phi}$ be a vector with only the intercept; we then have

$$\begin{aligned} \text{logit}\,(\mu_i) &= \beta_0 + \beta_1 \text{age} + \beta_2\ \text{sex} + \beta_3 \text{BMI}; \\ \log\,(\phi_i) &= \gamma_0; \\ i &= 1, \cdots, N, \end{aligned} \tag{6}$$

As in the previous case, we solve equation (4), in which μ_i and ϕ_i are replaced according to equation (5), using the `optim` function in `R`. Several examples are shown in the simulation studies presented later in this manuscript. In order to derive confidence intervals for the regression parameters, we use the asymptotic properties of MLEs for this regular parametric case. Letting $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$, we utilize the *observed information matrix* $\mathcal{J}(\boldsymbol{\theta})$, that is the negative of the second derivatives (the Hessian matrix) of the log-likelihood function

$$\left\{\mathcal{J}(\hat{\boldsymbol{\theta}})\right\}_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta}|\ \mathbf{y})\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{7}$$

which is a sample-based version of the Fisher information matrix. Mathematical formulas to compute the second derivatives of the ~~score function~~ log-likelihood are shown in Appendix A3. They involve polygamma functions, i.e. the derivatives of the logarithm of the gamma function, and several integrals without analytical solution. Therefore, approximation algorithms are required. Since the probability to make mistakes due to integral approximations or due to the numerical cancellation and/or to over-flow is high, we decided to approximate directly the Hessian matrix. For this reason, after having found the MLE, we adopt Richardson's extrapolation to compute accurate numerical second order derivative.

We have also investigated the Bayesian approach. We assigned non-informative (or little-informative) priors to each $\theta_g,\ g = 0, \ldots, p+q$, that is

$$\theta_g \sim N(0, \sigma^2 = 100)\ g = 0, \ldots, p+q, \tag{8}$$

normal distributions with large variance. We implemented the model in the Stan language. Stan is a new probabilistic programming language for Statistical inference written in C++. It is very flexible, since it allows to write directly the likelihood, and, using the Hamiltonian Monte Carlo algorithm for the MCMC and the BFGS algorithm for optimization, it is faster than other Bayesian interfaces, such as JAGS. We use the `R` package `RStan` [28] to call Stan and to estimate parameters. We also derive 95% credible intervals basing on posterior distributions. We obtained comparable results (simulations not shown) to the MLE version.

5 Fitting discretized beta distributions with random effects

Longitudinal studies yield multiple or repeated measurements on each individual. In order to take into account the correlations between ordinal repeated responses collected on the same subject, we follow a conditional approach as described in the Introduction and add random effects to the discretized beta models presented in the previous section. Mixed models have become very popular for the analysis of longitudinal data because they are flexible and widely applicable. They assume that measurements from a single subject share a set of latent, unobserved, random effects which are used to generate an association structure between the repeated measurements. Accordingly, let $y_{it} \in \{0, \ldots, n\}$, be the realization of the longitudinal ordinal variable Y_{it}, where $i \in \{1, ..., N\}$ refers to individuals and $t \in \{0, 1, ..., T_i\}$ to time; let $\mathbf{y}_i$ denote a $T_i \times 1$ vector of responses for the ith individual and assume that Y_{it}, given μ_{it} and ϕ_{it}, follows a discretized beta distribution, i.e. $(Y_{it}|\mu_{it}, \phi_{it}) \sim \text{Dbeta}(\mu_{it}, \phi_{it}, n)$. These kinds of models are also called two-stage models. In the first stage, we assume that

$$\begin{aligned} \text{logit}(\boldsymbol{\mu}_i) &= \mathbf{X}_i^{\mu}\boldsymbol{\beta} + \mathbf{Z}_i^{\mu}\mathbf{b}_{i\mu}; \\ \log(\boldsymbol{\phi}_i) &= \mathbf{X}_i^{\phi}\boldsymbol{\gamma} + \mathbf{Z}_i^{\phi}\mathbf{b}_{i\phi}; \\ i &= 1, \cdots, N; \end{aligned} \tag{9}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\phi}_i$ are the $T_i \times 1$ vectors of μ_{it} and ϕ_{it}, respectively. The population parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are treated as fixed effects, while $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ are random effects. Finally, $\mathbf{X}_i^{\mu}$ and $\mathbf{X}_i^{\phi}$ are design matrices of dimensions $T_i \times p$ and $T_i \times q$, respectively, linking the fixed effects to $\boldsymbol{\mu}_i$ and $\boldsymbol{\phi}_i$; $\mathbf{Z}_i^{\mu}$ and $\mathbf{Z}_i^{\phi}$ are the matrices of between-subject covariates, respectively, of dimensions $T_i \times k_{\mu}$ and $T_i \times k_{\phi}$, associated

to the random effects. If we suppose that the T_i values of $\mathbf{y}_i$, for the ith individual, are conditionally independent given $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$, then the likelihood for the ith individual is

$$\mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma} | \, \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) = \prod_{j=1}^{T_i} \left[I_{\frac{y_{ij}+1}{n+1}} \left(\mu_{ij}\phi_{ij}, (1-\mu_{ij})\phi_{ij}\right) - I_{\frac{y_{ij}}{n+1}} \left(\mu_{ij}\phi_{ij}, (1-\mu_{ij})\phi_{ij}\right) \right]. \quad (10)$$

At the second stage, we add assumptions on the random effects. In particular, in our model we assume $\mathbf{b}_{i\mu}$ follows a $\mathcal{N}(\mathbf{0}, \sigma_\mu^2 \mathbf{D}_\mu)$, and $\mathbf{b}_{i\phi}$ follows a $\mathcal{N}(0, \sigma_\phi^2 \mathbf{D}_\phi)$, independently of each other and satisfying the following condition

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i\mu} \\ \mathbf{b}_{i\phi} \end{pmatrix} \overset{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{D}_\mu & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_\phi \end{bmatrix} \right). \quad (11)$$

Here $\mathbf{D}_\mu$ are $k_\mu \times k_\mu$ and $\mathbf{D}_\phi$ are $k_\phi \times k_\phi$ positive-definite covariance matrices. Note we do not need balanced data and the model allows for explicit modelling and analysis of between- and within- individual variation.

Expanding the example in the previous section to mimick the kind of data appearing in our case study, let Y_{it} be the ith patient's response on a VAS for the nausea symptom at time t, with $t = 0, 1$, let $\mathbf{X}_i^\mu$ be the matrix in which each row is composed of the intercept, the value of age, an indicator variable for sex, the BMI (body mass index) and a time indicator and let $\mathbf{X}_i^\phi$ be a matrix with only the intercepts and time indicators; if we add only a random effect on μ, that is $b_i \sim \mathcal{N}(0, \sigma^2)$, we then have

$$\begin{aligned} \text{logit}\,(\mu_{it}) &= \beta_0 + \beta_1 \text{age} + \beta_2 \text{ sex} + \beta_3 \text{BMI} + \beta_4 \text{time} + b_i; \quad (12) \\ \log\,(\phi_{it}) &= \gamma_0 + \gamma_1 \text{time}; \\ & i = 1, \cdots, N; \quad t = 0, 1. \end{aligned}$$

Let $\boldsymbol{\theta}$ be the vector of variance and covariance parameters found in $\mathbf{D}_\mu$ and $\mathbf{D}_\phi$, with length $[k_\mu(k_\mu+1)/2 + k_\phi(k_\phi+1)/2]$. The classical approach is based on the maximum likelihood estimation of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ from the marginal distribution of $\mathbf{y}^T = (\mathbf{y}_1^T, ..., \mathbf{y}_N^T)$ (see among others Laird and Ware[29], Vaida and Xu[30], Stiratelli et al.[31]). Due to the chain rule and to the independence of $\mathbf{b}_{i\mu}$ and $\mathbf{b}_{i\phi}$ we have

$$f_{(\mathbf{Y}_i, \mathbf{b}_i)}(\mathbf{y}_i, \mathbf{b}_i) = P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_i}(\mathbf{b}_i) = P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}). \quad (13)$$

Therefore, the marginal distribution of $\mathbf{y}_i$ is given by

$$P(\mathbf{Y}_i = \mathbf{y}_i) = \int_{R^\mu} \int_{R^\phi} P(\mathbf{Y}_i = \mathbf{y}_i | \mathbf{b}_i) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi}) \, d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi}. \quad (14)$$

If we assume that observations on different individuals are independent, we obtain that the marginal log-likelihood of the data can be written in the following way

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{N} \log \int_{R^{\mu}} \int_{R^{\phi}} \mathcal{L}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}|\, \mathbf{y}_i, \mathbf{b}_{i\mu}, \mathbf{b}_{i\phi}) f_{\mathbf{b}_{i\mu}}(\mathbf{b}_{i\mu}) f_{\mathbf{b}_{i\phi}}(\mathbf{b}_{i\phi})\, d\mathbf{b}_{i\mu} d\mathbf{b}_{i\phi}. \tag{15}$$

There are computational difficulties in maximizing the marginal likelihood with the MLE approach. The EM algorithm is very slow to converge and derivatives are not automatically computed. On the other hand, usual integral approximations, such as the Laplace's approximation or adaptive Gaussian approximation (see, e.g., the R package lme4 or the SAS proc nlmixed), can be used with acceptable results when the number of random effects is not high (usually not more than three). Moreover, as for the random effects, modelling their variance-covariance matrices is difficult in the general case.

Instead of doing so, we go directly for the Bayesian approach. We rewrite the model as a Bayesian hierarchical model, choosing little-informative priors on the regression parameter and a non informative Wishart distribution for the inverse variance-covariance matrices. Other priors on the variance-covariance matrices could be taken [32;33] . For this purpose, we first run the model without random effects, as described in the previous section then, within a sort of empirical Bayes approach, we use the estimated parameters as means for normal prior distributions of the regression fixed-effect parameters, assigning them at the same time large prior variances. Therefore, the final Bayesian hierarchical model is the following

$$\begin{aligned}
y_{ij} &\sim \text{Dbeta}(\mu_{ij}, \phi_{ij}, n) \\
\text{logit}\,(\mu_{ij}) &= \mathbf{x}_{ij}^{\mu}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mu}\mathbf{b}_{i\mu} \\
\log\,(\phi_{ij}) &= \mathbf{x}_{ij}^{\phi}\boldsymbol{\gamma} + \mathbf{z}_{ij}^{\phi}\mathbf{b}_{i\phi} \\
\mathbf{b}_{i\mu} &\sim N(\mathbf{0}, \mathbf{D}_{\mu}) \\
\mathbf{b}_{i\phi} &\sim N(\mathbf{0}, \mathbf{D}_{\phi}) \\
[\boldsymbol{\beta}, \boldsymbol{\gamma}] &\sim N([\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*], \sigma^{2*}\mathbf{I}) \\
\mathbf{D}_{\mu}^{-\mathbf{1}} &\sim W(\mathbf{V}_1, n_1) \\
\mathbf{D}_{\phi}^{-\mathbf{1}} &\sim W(\mathbf{V}_2, n_2)
\end{aligned} \tag{16}$$

where ~~$\boldsymbol{\theta}^*$ is~~ $[\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*]$ are computed using the Empirical Bayes method described in the previous paragraph, and σ^*, $\mathbf{V}_1$, $\mathbf{V}_2$, n_1 and n_2 are parameters of the prior distributions.

In our specific application, we will set σ^* equal to 100 and we will use a single scalar random effect for μ, as appearing in Eq.12, and assign to its standard deviation an inverse gamma distribution, the scalar version of the inverse Wishart distribution, with parameters 0.01 and 0.01 as prior.

Increasing the sample size and the number of regression parameters could lead to difficulties in achieving estimation convergence. For this reason, the estimation should be run several times adapting sampling parameters (as suggested by software, for example).

6 Simulation studies - a practical overview

In this section we evaluated first the ability of the regression model to estimate both μ and ϕ in absence of covariates. As simulation values we chose $\mu \in (0.15, 0.3, 0.5, 0.7, 0.85)$, $\phi \in (1, 3, 5, 7, 9, 11)$ and n varying from 3 to 15. For each combination of these three parameters, we run 1000 trials in which we simulated 100 observations. Then we run the regression model in each trial and we computed the median, the first and the third quantile of estimates. Detailed tables are summarized in the Supplementary Material. As we can see, μ is better estimated than ϕ, which has a larger uncertainty. Only with $n = 3$, μ near to the border of its support ad high ϕ, our method is not able to estimate the parameters acceptably. The problem disappears when the number of observations increases (results not shown).

In the second batch of our simulation study we added covariates to μ and ϕ. We simulated 5 different covariates: x_1 normally distributed with mean equals to 50 and σ equals to 10, x_2 from a Poisson distribution with rate equals to 10 and mean shifted by 10, x_3 from a Bernoulli distribution with probability of success equals to 0.5, x_4 from a normal distribution with mean equals to 10 and σ equals to 1 and x_5 from a uniform distribution. Then, we used those simulated values to create two scenarios. In the first scenario, we imposed 4 covariates to μ and 2 to ϕ. In details,

$$\begin{aligned} \text{logit}\,(\mu) &= 1 + 0.2x_1 - 0.3x_2 + x_3 - 0.5x_4; \\ \log\,(\phi) &= 3.5 + x_3 - 0.2x_4. \end{aligned}$$

1000 trial with 100 patients and $n = 5$ were simulated. Estimation results are shown in Figure 1, left panel. In the second we imposed 5 covariate to μ and 3 to ϕ:

$$\begin{aligned} \text{logit}\,(\mu) &= 1 + 0.2x_1 - 0.3x_2 + x_3 - 0.6x_4 + 6x_5; \\ \log\,(\phi) &= 3.5 + x_3 - 0.3x_4 + 3x_5; \end{aligned}$$

and results are summarized in Figure 1, right panel. Boxplots of estimation results are plotted along with the real value of the parameters, represented by triangular black points. As we expected from the previous results, estimation of $\boldsymbol{\beta}$ is more precise than the one of $\boldsymbol{\gamma}$. In any case, the regression gives good

results. Unifying these results to the ones achieved in the previous study, we suggest not to add too many regressors on ϕ. It is natural to add more covariates on the mean, since in many cases it contains the most important information. Instead, it could be useful to let ϕ depend on specific variables, such as sex, to point out differences between groups. An example is given at the end of the section.

[Figure 1 about here.]

The third step regards the random effects. For the sake of clarity we added only one random effect to μ, in order to be closer to the model which we used in the real gastrointestinal example. We set up two scenarios with 15 subjects, 5 measurements for each subject, $\beta_0 = -0.405$ and $\gamma_0 = 1.061$. In the fist scenario, the standard deviation of the random effect is equals to 0.7 and in the second to 2. Results are shown in Figure 2, in a similar way to Figure 1.

[Figure 2 about here.]

We can see that the estimate of the standard deviation of random effects becomes less precise as its true value increase. This is not unexpected due to the small number of subject and the behaviour of the link functions. With the logit link we do not expect high covariate effects, since values in the tails of the function are not enough distinguishable.

Finally, we compared our model to the cumulative logit model (`ordinal` package[34]), the most used model in the field of ordinal regression. Again, we present two situations. We suppose to have 50 males and 50 females answering to a question about some preference on a 0 to 5 scale. In the first example, male answers follow a discrete uniform distribution, which means that males are hesitant. Instead, females have more focused responses and their probabilities to chose the number from 0 to 5 are $p_0 = 0$, $p_1 = 0.1$, $p_2 = 0.8$, $p_3 = 0.1$, $p_4 = 0$ and $p_5 = 0$. In the second example, the two groups have opposite trends and $p_0 = 0.3$, $p_1 = 0.45$, $p_2 = 0.2$, $p_3 = 0.05$, $p_4 = 0$ and $p_5 = 0$ are assigned to males. Instead females answer following $p_0 = 0$, $p_1 = 0$, $p_2 = 0.05$, $p_3 = 0.2$, $p_4 = 0.45$ and $p_5 = 0.3$. We simulated 1000 trials for each example and then we compared the ability to estimate the probability associated to each category. Results are shown in Figure 3.

[Figure 3 about here.]

As we can see, in the first example CLM is not able to achieve good results. The proportional odds assumption is violated. In that case, a CLM with category dependent parameters is required, increasing model complexity. On the other hand, Dbeta can well estimate all the probabilities and β_{sex} and γ_{sex} were significant, respectively, at 48% and at 98% of the times. We need to remind that the difference in mean between males and female is not high (2 vs 2.5), whereas the precision is very different. Dbeta

model is able to point out these facts. In the second example, both models are able to well estimate the probabilities. Moreover, Dbeta suggested all the times that mean is significant. Those examples explaine the easy interpretation of Dbeta regression results and its intuitive application. CLM has more flexibility, due to the high number of parameters, and can be run also in the case in which the answer follow a multimodal distribution. However its interpretation is not easy, since the parameters express the probabilities to pass to higher or lower values. A plot is usually needed to understand the trend. On the other hand, the information given by Dbeta model are very intuitive and ready to use, and the model can be applied in unimodal settings, the most common ones.

7 Results

The pre/post study is a first example of a common medical research design in which a single baseline health status measurement is obtained, an intervention is administered, then a single follow-up measurement is collected. The design can then be viewed as the simplest form of a prospective longitudinal study.

In the dataset presented in Section 2, we do not have a pre/post study with a therapeutic intervention, but a similar example of a simple longitudinal design. The WLT was administered to two separate groups, healthy volunteers and GERD patients, then followed up at two time points to gauge differences between the two groups while keeping into account the natural course in time of gastric sensations after filling the stomach with water. The main goal was to investigate if there is any difference in gastric sensations between healthy volunteers and GERD patients as assessed by the WLT. In a sense, the study was more of a clinical epidemiology diagnostic experiment than a clinical trial.

To do so, we applied the discretized beta model with random effects to each of the three symptoms representing gastric sensations: nausea, postprandial fullness and bloating. For each symptom, the simplest model with a single random effect on the intercept of μ was chosen; to which we added a series of fixed effect parameters to model the effects of covariates.

In Figure 4, 95% equal tail posterior credible intervals are shown for the fixed effect parameters relative to μ and ϕ and for σ. For the sake of brevity, we only show the models with the highest number of statistically significant estimated coefficients (meaning zero is outside of the credible interval) and highest posterior average loglikelihood (as in the standard `rstan` output).

As we can see in Figure 1, panel (a), the nausea VAS score at time 1 is significantly lower than at time 0 ($\hat{\beta}_{time1} = -1.85$, CI $(-2.06, -1.65)$). A higher BMI ($\hat{\beta}_{BMI} = -0.11$, CI $(-0.21, -0.02)$) leads to a lower average score and GERD patients ($\hat{\beta}_{status1} = 1.83$, CI $(1.33, 2.33)$) have on average higher scores than controls. Water intake during WLT is not significant. Variability in the score is higher at time 1 ($\hat{\gamma}_{time1} = 1.39$, CI $(0.71, 2.19)$) and lower for GERD patients ($\hat{\gamma}_{status1} = -2.02$, CI $(-2.76, -1.28)$).

Similar considerations apply to the postprandial fullness symptom (Figure 4, panel (b)) about the time predictor, since $\hat{\beta}_{time1} = -2.30$ with CI $(-2.59, -2.01)$ and $\hat{\gamma}_{time1} = 1.85$, CI $(0.90, 3.32)$. A higher volume of water intake leads to lower VAS score ($\hat{\beta}_{water\_intake} = -1.40$, CI $(-2.66, -0.96)$) while disease status is not significant.

About bloating (Figure 4, panel (c)), time acts in the same way as for the previous symptoms, $\hat{\beta}_{time1} = -2.11$ with CI $(-2.40, -1.82)$ and $\hat{\gamma}_{time1} = 1.13$, CI $(0.46, 1.95)$. As for nausea, GERD patients ($\hat{\beta}_{status1} = 0.80$, CI $(0.40, 1.21)$) have on average higher scores than controls.

[Figure 4 about here.]

8 Discussion

In this work we propose a new discrete probability distribution useful to analyse ordered categorical data. The discretized beta distribution has an highly flexible shape and it can be either over-dispersed or under-dispersed with respect to the binomial distribution. It has only two parameters, μ and ϕ, which have a very clear interpretation: μ is the mean of the beta latent variable and the greater its value the higher the probability to obtain a higher score; ϕ is a precision parameter of the latent variable and the higher its value the lower the variance. For an ordinal variable expressing a subjective evaluation, ϕ can be interpreted as a degree of confidence of such evaluation.

We have shown some practical examples in order to clarify the easy interpretation and application of this new model. We suggest not to add too many covariate on the precision parameters, since the estimations are more precise on the mean. On the other hand, it is natural to think that covariate should have more impact on the mean than on precision. In case of small samples, we will not suggest to apply Dbeta if the number of possibilities is smaller than 4, since in that case the precision parameter could be misleading. On the other hand, this model permits a global regression associated with intuitive result descriptions, as in the example of male and female answers. This is a clear advantage over the CLM, which in this case should be either hyper-parametrized or applied separately for males and females.

We have applied our model to a real data set, described in Section 2, obtaining interesting results. We proved that there are differences in nausea and bloating average VAS scores between GERD patients and healthy volunteers, while the volume of water recorded is statistically significant only for the postprandial fullness VAS score. These results are useful for physicians when it comes to evaluating the usefulness of the WLT in diagnosing GERD patients.

Acknowledgements

We thank Edda Battaglia from the gastroenterology department of Hospital Cardinal Massaia, Asti, who provided the data used in this research and the two anonymous reviewers for their constructive comments. We would also like to show our gratitude to the InSPiRe project.

Conflict of Interest

The authors have declared no conflict of interest.

References

1. Agresti A. *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons, 2010.
2. Becker M. *Encyclopedia of Biostatistics*, volume 6, chapter Ordered categorical data. John Wiley, 2005. pp. 3869–3876.
3. Snell E. A scaling procedure for ordered categorical data. *Biometrics* 1964; 20: 592–607.
4. Bock R and Jones L. *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day, 1968.
5. Cullagh M. Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society: Series B* 1980; 42: 109–142.
6. Goodman LA. Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* 1979; 74(367): 537–552.
7. Muniz-Terrera G, van den Hout A, Rigby R et al. Analysing cognitive test data: Distributions and non-parametric random effects. *Statistical methods in medical research* 2012; DOI:10.1177/0962280212465500.
8. Stasinopoulos DM and Rigby RA. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software* 2007; 23: 1–46. URL http://www.jstatsoft.org/v23/i07.
9. D'Elia A and Piccolo D. A mixture model for preferences data analysis. *Computational Statistics & Data Analysis* 2005; 49(3): 917–934.
10. Iannario M. Hierarchical cub models for ordinal variables. *Communications in Statistics-Theory and Methods* 2012; 41(16-17): 3110–3125.
11. Lee K and Daniels MJ. Marginalized models for longitudinal ordinal data with application to quality of life studies. *Statistics in Medicine* 2008; 27(21): 4359–4380.
12. Crouchley R. A random-effects model for ordered categorical data. *Journal of the American Statistical Association* 1995; 90(430): 489–498.
13. Thomas R and Have T. A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* 1996; 52: 473–491.
14. Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; 88(421): 9–25.

15. Engel B. A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* 1998; 40(2): 141–154.
16. Gajewski BJ, Hart S, Bergquist-Beringer S et al. Inter-rater reliability of pressure ulcer staging: ordinal probit bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine* 2007; 26(25): 4602–4618.
17. Tan M, Qu Y, Mascha E et al. A bayesian hierarchical model for multi-level repeated ordinal data: analysis of oral practice examinations in a large anaesthesiology training programme. *Statistics in Medicine* 1999; 18(15): 1983–1992.
18. Lunn DJ, Wakefield J and Racine-Poon A. Cumulative logit models for ordinal data: a case study involving allergic rhinitis severity scores. *Statistics in Medicine* 2001; 20(15): 2261–2285.
19. Lipsitz SR, Kim K and Zhao L. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* 1994; 13(11): 1149–1163.
20. Bonney GE. Logistic regression for dependent binary observations. *Biometrics* 1987; 43: 951–973.
21. Muenz LR and Rubinstein LV. Markov models for covariate dependence of binary sequences. *Biometrics* 1985; 41: 91–101.
22. Battaglia E, Grassini M, Navino M et al. Water load test before and after ppi therapy in patients with gastro-oesophageal reflux disease. *Digestive and Liver Disease* 2007; 39(12): 1052–1056.
23. Johnson VE and Albert JH. *Ordinal data modeling.* Statistics for Social Science and Public Policy. New York, NY: Springer. 258 p. , 1999. DOI:10.1007/b98832.
24. Ferrari S and Cribari-Neto F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 2004; 31(7): 799–815.
25. Johnson N and Kotz S. *Distributions in statistics*. Wiley series in probability and mathematical statistics, Wiley, 1970. ISBN 9780471443605.
26. Punzo A. Discrete beta-type models. In *Classification as a Tool for Research*. Springer, 2010. pp. 253–261.
27. Byrd RH, Lu P, Nocedal J et al. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* 1995; 16(5): 1190–1208.
28. Team SD. Rstan: the r interface to stan, version 2.8.0, 2015. URL http://mc-stan.org/rstan.html.
29. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38: 963–974.
30. Vaida F and Xu R. Proportional hazards model with random effects. *Statistics in Medicine* 2000; 19(24): 3309–3324.
31. Stiratelli R, Laird N and Ware JH. Random-effects models for serial observations with binary response. *Biometrics* 1984; 40: 961–971.
32. Daniels MJ. A prior for the variance in hierarchical models. *Canadian Journal of Statistics* 1999; 27(3): 567–578.

33. Chung Y, Gelman A, Rabe-Hesketh S et al. Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics* 2015; 40(2): 136–157.
34. Christensen RHB. ordinal—regression models for ordinal data, 2015. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.
35. Dennis Jr JE and Schnabel RB. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam, 1996.

Appendix

A.1. Identifiability of the model

We say that a model is identifiable if it is theoretically possible to learn the true values of parameters underlying the model after obtaining an infinite number of observations from it. Mathematically, a parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability distributions. Identifiability is a property of the model, not of an estimator or estimation procedure. If the model is not identifiable, there are some difficulties in doing inference. In the following, we prove that the discretized beta family is identifiable.

Let $n = 2$. Suppose there exist $(\mu_1, \phi_1) \neq (\mu_2, \phi_2)$ such that

$$P(Y_1 = 0) = P(Y_2 = 0);\ P(Y_1 = 1) = P(Y_2 = 1);\ P(Y_1 = 2) = P(Y_2 = 2) \tag{17}$$

where $Y_1 \sim \text{Dbeta}(\mu_1, \phi_1)$ and $Y_2 \sim \text{Dbeta}(\mu_2, \phi_2)$. For the sake for clarity, let $a_1 = \mu_1\phi_1$, $b_1 = (1 - \mu_1)\phi_1$, $a_2 = \mu_2\phi_2$ and $b_2 = (1 - \mu_2)\phi_2$. From eq. (17) we obtain that

$$I_{\frac{1}{3}}(a_1, b_1) = I_{\frac{1}{3}}(a_2, b_2) \quad \text{and} \quad I_{\frac{2}{3}}(a_1, b_1) = I_{\frac{2}{3}}(a_2, b_2). \tag{18}$$

Equation (18) implies that the densities of the latent variables have at least 3 intersections: at least one between $\left(0, \frac{1}{3}\right)$, at least one between $\left(\frac{1}{3}, \frac{2}{3}\right)$ and at least one between $\left(\frac{2}{3}, 1\right)$. However, it is not possible that two different beta densities have more than two intersections over $(0, 1)$. Thus, we have that the distributions coincide, $a_1 = a_2$ and $b_1 = b_2$, which implies, $\mu_1 = \mu_2$ and $\phi_1 = \phi_2$. The discretized beta model is identifiable. The result can be easily extended to $n > 2$ repeating the same argumentat on

$$I_{\frac{1}{n+1}}(a_1, b_1) = I_{\frac{1}{n+1}}(a_2, b_2) \quad \text{and} \quad I_{\frac{n}{n+1}}(a_1, b_1) = I_{\frac{n}{n+1}}(a_2, b_2).$$

A.2. Over-dispersion and Under-dispersion

In this appendix, we compare our discretized beta distributions with other two discrete distributions, the binomial and the beta-binomial, defined on the same support $[0, 1, \ldots, n]$. Let X have a binomial distribution with parameters n and p. Its probability mass function (pmf) is defined as

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, \cdots, n; \tag{19}$$

where $p \in [0,1]$ is called the probability of success. Let Y be a random variable with a beta-binomial density with parameters n, a and b, that is

$$P(Y=k) = \binom{n}{k} \frac{B(a+x, b+n-x)}{B(a,b)}, \quad k = 0, \cdots, n; \tag{20}$$

where $B(p,q)$ is the beta function and a and b are two positive parameters. Thus we have that $E[X] = np$, $\text{Var}[X] = np(1-p)$, $E[Y] = n\frac{a}{a+b}$ and $\text{Var}[Y] = n\frac{ab}{(a+b)^2}\frac{a+b+n}{a+b+1}$. It is known that, if X and Y have the same expected value, i.e. $p = \frac{a}{a+b}$, the variance of Y is greater than the variance of X for all values of a and b. Due to this reason, the beta-binomial model is a popular and analytically tractable alternative to the binomial that captures *overdispersion* with respect to the binomial model.

Let Z to be a random variable with the discretized beta density (2). Then, $E[Z] = n - \sum_{i=1}^{n} I_{\frac{i}{n+1}}(\mu\phi, (1-\mu)\phi)$ and $\text{Var}[Z] = \sum_{i=1}^{n}(2n-2i+1) I_{\frac{i}{n+1}}(\mu\phi,(1-\mu)\phi) - \left(\sum_{i=1}^{n} I_{\frac{i}{n+1}}(\mu\phi,(1-\mu)\phi)\right)^2$.

[Figure 5 about here.]

Let $E[Z] = \mu' = np$. We can write $\text{Var}[Z]$ in the following way

$$\text{Var}[Z] = (n-\mu')(n+\mu'+1) - 2\sum_{i=1}^{n} iI_i, \tag{21}$$

where $I_i = I_{\frac{i}{n+1}}(\mu\phi, (1-\mu)\phi)$ and (μ, ϕ) is a parameter pair which satisfy $E[Z] = \mu'$. We have the minimum variance when the mass is, in the limit case, concentrated in a single point of the set $[0, 1, \ldots, n]$ and the point is equal to μ' (a degenerate mass function). It is obtained by discretizing, for example, a beta distribution with $\mu = \frac{\mu'}{n+1} + \frac{1}{2}\frac{1}{n+1}$ and $\phi \to +\infty$ (the limit case where the variance of the beta latent variable tends to zero). In this case, we have

$$2\sum_{i=1}^{n} iI_i \to 2\sum_{i=\mu'+1}^{n} i \cdot 1 = n^2 + n - \mu'^2 - \mu'$$

and eq. (21) tends to zero.

On the other hand, discretized beta distributions can have U-shapes. For a fixed μ', every (μ, ϕ), such that $E[Z] = \mu'$ and $\phi < \min\left(\frac{1}{\mu}, \frac{1}{1-\mu}\right)$*, generates a U-shape distribution (it follows from the properties of the beta latent distribution[25]). Thus, it can be overdispersed. Since the variance is a continuous function of μ and ϕ, and $E[Z] = \mu'$ is also a continuous curve†, we have that discretized beta distribution can be underdispersed, overdispersed or it can have the same variance of a binomial with the same mean value.

Thus, if we set $E[Z] = \mu'$ and $\text{Var}[Z] = \sigma^{2'}$, where μ' and $\sigma^{2'}$ are admissible values, in order to find μ and ϕ we have to solve the following system

$$\begin{cases} n - \sum_{i=1}^{n} I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) = \mu' \\ (n-\mu')\left(n+\mu'+1\right) - 2\sum_{i=1}^{n} i I_{\frac{i}{n+1}}\left(\mu\phi, (1-\mu)\phi\right) = \sigma^{2'}. \end{cases} \tag{22}$$

This is a non-linear system and we have not found a closed form solution, so numerical algorithms such as the Broyden Secant method[35], must be used. It is also possible to solve the minimization problem linked to this system

$$\min_{\mu,\phi}\left[\left(E[Z]-\mu'\right)^2 + \left(\text{Var}[Z] - \sigma^{2'}\right)^2\right], \tag{23}$$

with, for example, a quasi-Newton method.

[Figure 6 about here.]

A.3. Computation of the second derivatives for the observed information matrix

For the sake of clarity, we will use the following notation

$$h(x, \mu, \phi) = x^{\mu\phi-1}(1-x)^{(1-\mu)\phi-1}, \tag{24}$$

*The existence of such points can be proved graphically intersecting the surface plotted in Figure 5 with the plane $z = \mu'$ in the domain

$$D(\mu, \phi) = \begin{cases} \phi < \frac{1}{1-\mu}, & \text{if } 0 < \mu \leq \frac{1}{2} \\ \phi < \frac{1}{\mu}, & \text{if } \frac{1}{2} < \mu < 1 \end{cases}$$

†This curve is the intersection between the surface plotted in Figure 5 and the plane $z = \mu'$.

and we compute the first derivatives as

$$\frac{\partial}{\partial \mu}\log P(Y=k) = -\phi\psi(\mu\phi) + \phi\psi((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx} \tag{25}$$

$$\frac{\partial}{\partial \phi}\log P(Y=k) = \psi(\phi) - \mu\psi(\mu\phi) - (1-\mu)\psi((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx} \tag{26}$$

where $\psi(t)$ is the *digamma* function, that is defined as the logarithmic derivative of the gamma function

$$\psi(t) = \frac{d}{dt}\log\Gamma(t). \tag{27}$$

Thus, the second derivatives can be written as

$$\frac{\partial^2}{\partial \mu^2}\log P(Y=k) = -\phi^2\psi_1(\mu\phi) + \phi^2\psi_1((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi^2\log^2\left(\frac{x}{1-x}\right)\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx} + \\ - \left(\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\,dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\,dx}\right)^2 \tag{28}$$

$$\frac{\partial^2}{\partial \phi^2}\log P(Y=k) = \psi_1(\phi) - \mu^2\psi_1(\mu\phi) - (1-\mu)^2\psi_1((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]^2\, dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\, dx} + \\ - \left(\frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\, dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\, dx}\right)^2 \tag{29}$$

$$\frac{\partial^2}{\partial \phi \partial \mu}\log P(Y=k) = \psi_1(\mu\phi) - \mu\phi\psi_1(\mu\phi) + \psi_1((1-\mu)\phi) + (1-\mu)\phi\psi_1((1-\mu)\phi) + \\ + \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\phi\log\left(\frac{x}{1-x}\right)\, dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\, dx} + \\ - \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\left[\mu\log x + (1-\mu)\log(1-x)\right]\, dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\, dx} * \\ * \frac{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\phi\log\left(\frac{x}{1-x}\right)\, dx}{\int_{\frac{k}{n+1}}^{\frac{k+1}{n+1}} h(x,\mu,\phi)\, dx} \tag{30}$$

where $\psi_1(t)$ is the *trigamma* function defined as

$$\psi_1(t) = \frac{d^2}{d^2 t}\log\Gamma(t). \tag{31}$$

In the case of covariates, we can compute the first derivatives using the chain rule derivation

$$\frac{\partial f}{\partial \beta_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c} + \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \beta_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \beta_c}, \quad c = 1, \cdots, p; \tag{32}$$

$$\frac{\partial f}{\partial \gamma_c} = \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial \gamma_c} + \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \gamma_c} = \frac{\partial f}{\partial \phi}\frac{\partial \phi}{\partial \gamma_c}, \quad c = 1, \cdots, q. \tag{33}$$

and the second derivatives as

$$\begin{aligned}\frac{\partial^2 l}{\partial\beta_d\partial\beta_c} &= \frac{\partial}{\partial\beta_d}\left(\frac{\partial f}{\partial\beta_c}\right) = \frac{\partial}{\partial\beta_d}\left(\frac{\partial f}{\partial\mu}\frac{\partial\mu}{\partial\beta_c}\right)\\ &= \frac{\partial^2 l}{\partial\mu^2}\frac{\partial\mu}{\partial\beta_d}\frac{\partial\mu}{\partial\beta_c} + \frac{\partial f}{\partial\mu}\frac{\partial^2\mu}{\partial\beta_d\partial\beta_c}, \quad c=1,\cdots,p;\ d=1,\cdots,p;\end{aligned} \tag{34}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial\gamma_d\partial\gamma_c} &= \frac{\partial}{\partial\gamma_d}\left(\frac{\partial f}{\partial\gamma_c}\right) = \frac{\partial}{\partial\gamma_d}\left(\frac{\partial f}{\partial\phi}\frac{\partial\phi}{\partial\gamma_c}\right)\\ &= \frac{\partial^2 l}{\partial\phi^2}\frac{\partial\phi}{\partial\gamma_d}\frac{\partial\phi}{\partial\gamma_c} + \frac{\partial f}{\partial\phi}\frac{\partial^2\phi}{\partial\gamma_d\partial\gamma_c}, \quad c=1,\cdots,q;\ d=1,\cdots,q;\end{aligned} \tag{35}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial\gamma_d\partial\beta_c} &= \frac{\partial}{\partial\gamma_d}\left(\frac{\partial f}{\partial\beta_c}\right) = \frac{\partial}{\partial\gamma_d}\left(\frac{\partial f}{\partial\mu}\frac{\partial\mu}{\partial\beta_c}\right)\\ &= \frac{\partial^2 l}{\partial\phi\partial\mu}\frac{\partial\phi}{\partial\gamma_d}\frac{\partial\mu}{\partial\beta_c} + \frac{\partial f}{\partial\mu}\frac{\partial^2\mu}{\partial\gamma_d\partial\beta_c}\\ &= \frac{\partial^2 l}{\partial\phi\partial\mu}\frac{\partial\phi}{\partial\gamma_d}\frac{\partial\mu}{\partial\beta_c} \qquad c=1,\cdots,p;\ d=1,\cdots,q.\end{aligned} \tag{36}$$

The derivatives with respect to μ and ϕ are the same written in the previous paragraphs. We also have that

$$\frac{\partial\mu}{\partial\beta_c} = \frac{x_c}{2\left(1+\cosh(\mathbf{x}\boldsymbol{\beta})\right)}, \quad c=1,\cdots,p; \tag{37}$$

$$\frac{\partial^2\mu}{\partial\beta_d\partial\beta_c} = \frac{-x_d x_c\sinh(\mathbf{x}\boldsymbol{\beta})}{4\left(1+\cosh(\mathbf{x}\boldsymbol{\beta})\right)^2}, \quad c=1,\cdots,p; \quad d=1,\cdots,p; \tag{38}$$

$$\frac{\partial\phi}{\partial\gamma_c} = x_c e^{\mathbf{x}\boldsymbol{\gamma}}, \quad c=1,\cdots,q; \tag{39}$$

$$\frac{\partial^2\phi}{\partial\gamma_d\partial\gamma_c} = x_d x_c e^{\mathbf{x}\boldsymbol{\gamma}}, \quad c=1,\cdots,q; \quad d=1,\cdots,q; \tag{40}$$

where x_l denote the value of the vector $\mathbf{x}$ at the position l.

List of Figures

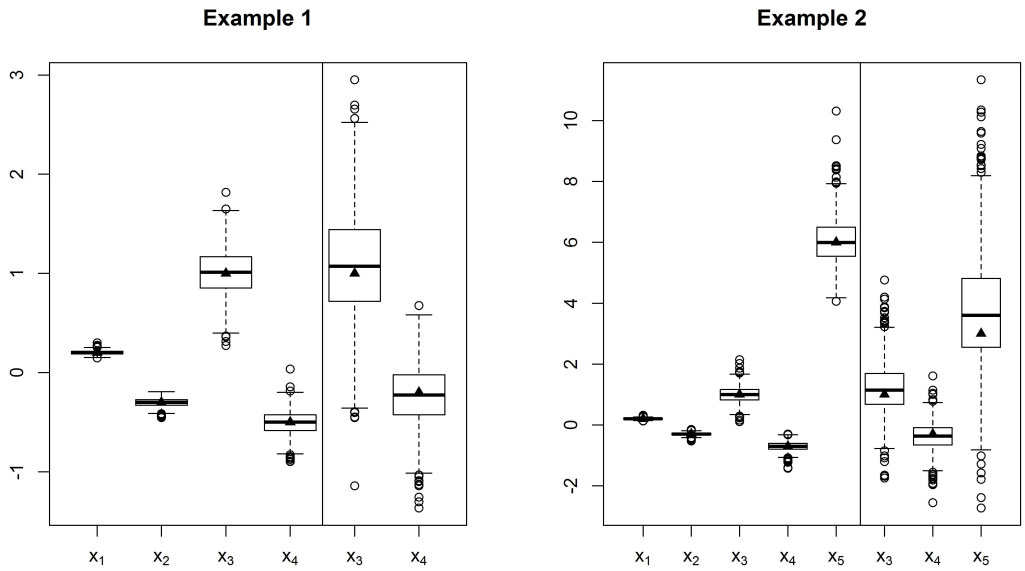


Figure 1. Boxplots of the sampling distribution of the estimated parameters over 1000 trials with covariates. Black triangles represent the true value of the regression parameters related to x_i. On the left-hand side of the vertical line, the results for μ are plotted, on the right-hand side the ones for ϕ. The covariates x_i are described in the text.

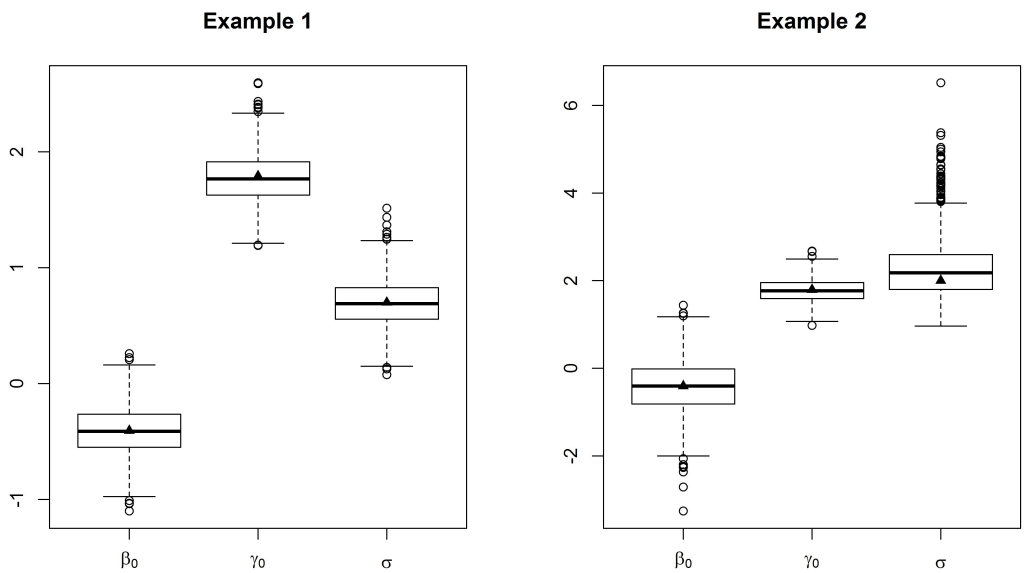


Figure 2. Boxplots of the sampling distributions of the estimated parameters over 1000 trials in the case of random effects. Black triangles represent the true value of the regression parameters.

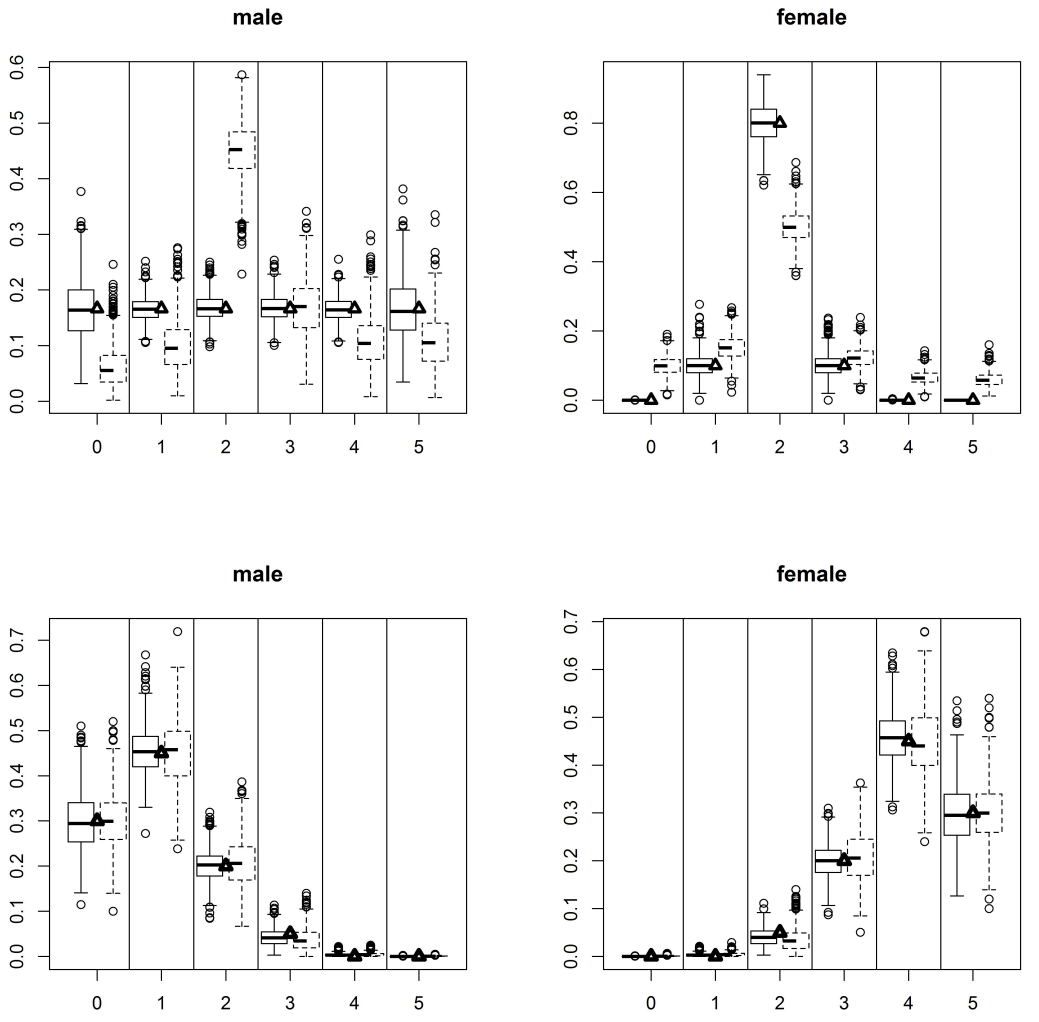
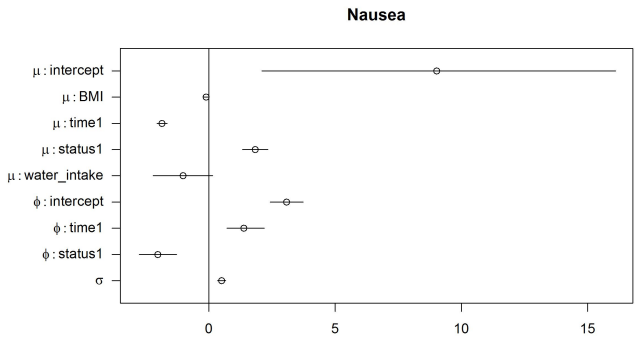
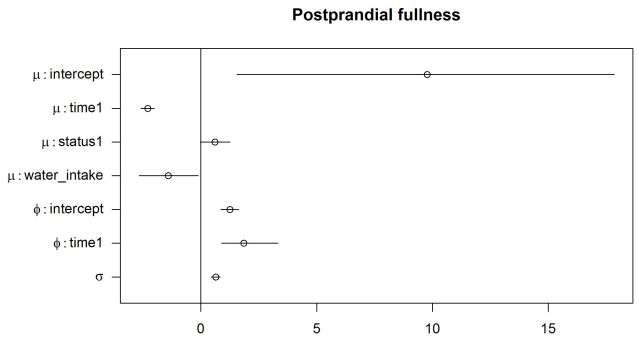


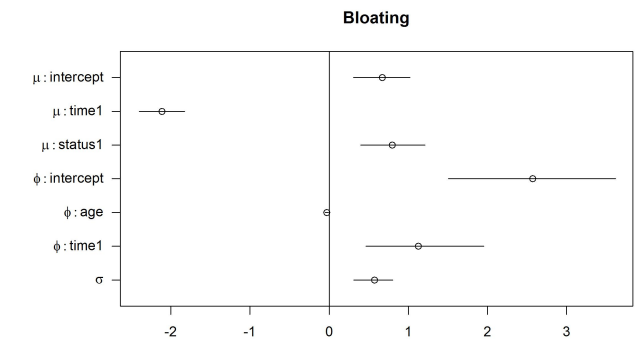
Figure 3. Estimated probability for each class. Black triangles represent the true value of probability. Solid line refers to Dbeta method and dashed to CLM regression. On the top: plots show the results of the first example. On the bottom: plots show the results of the second example.



(a)



(b)



(c)

Figure 4. 95% equal tail posterior credible intervals for sigma and the fixed-effect regression coefficients of time (i.e. time 1 vs. time 0, so that this effect is naturally negative), BMI (body mass index), status (GERD patients 1, vs. healthy volunteers 0), water intake (logarithms, in [ml]), age (in years), for the symptoms nausea (a), postprandial fullness (b) and bloating (c).

Prepared using sagej.cls

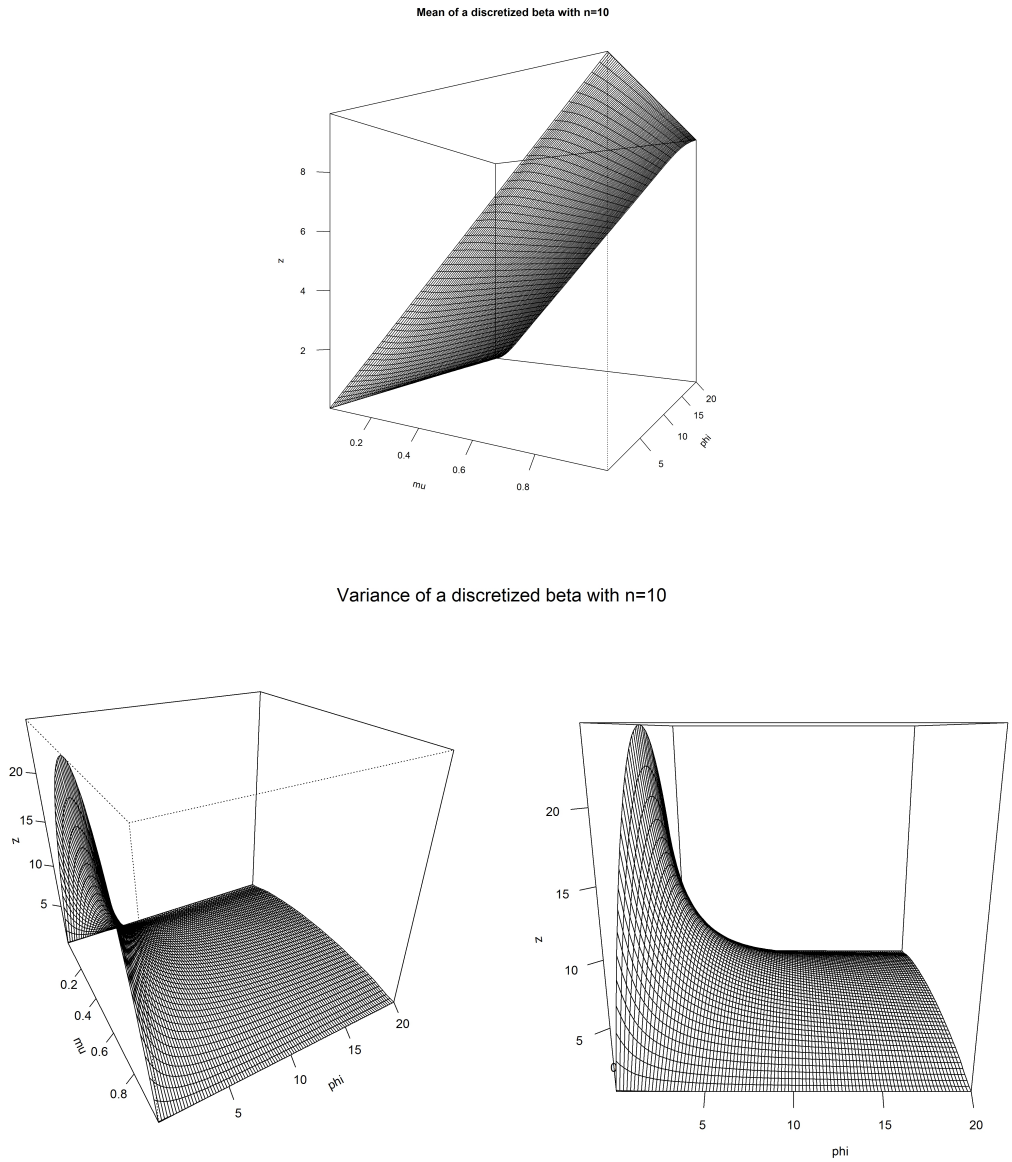


Figure 5. Top: expectation of a discretized beta distribution with $n = 10$, plotted for $(\mu, \phi) \in (0, 1) \times (0, 20]$. Bottom: variance of a discretized beta distribution with $n = 10$, plotted for $(\mu, \phi) \in (0, 1) \times (0, 20]$. Two different views of the same plot are shown.

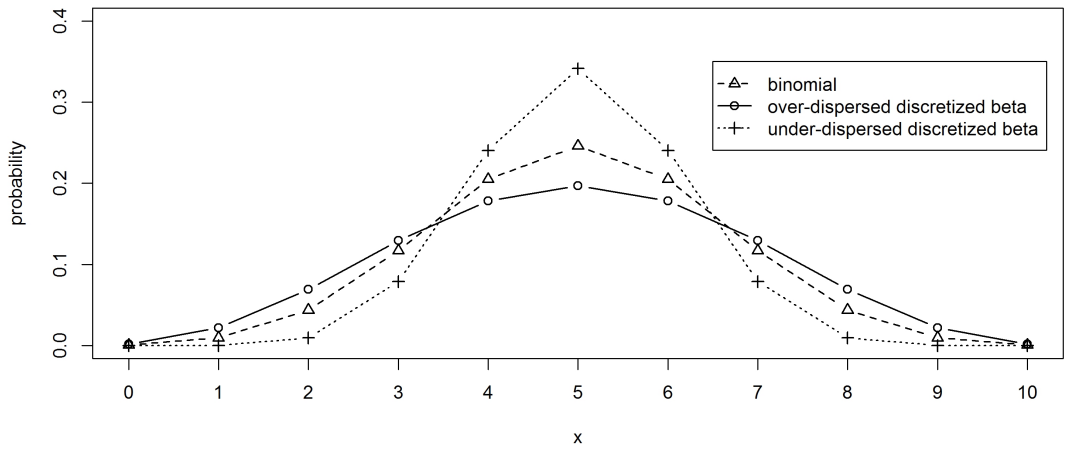


Figure 6. A binomial distribution is plotted together with both under-dispersed and over-dispersed discretized beta distributions with the same expected value (equal to 5).